Joris van de Klundert, Laurens Wormer

ASAP: The After Salesman Problem

RM/08/054

JEL code: C61

Maastricht research school of **E**conomics
of **TE**chnology and **OR**ganizations

Universiteit Maastricht
Faculty of Economics and Business Administration
P.O. Box 616
NL - 6200 MD Maastricht

phone : ++31 43 388 3830
fax : ++31 43 388 4873

ASAP: THE AFTER SALESMAN PROBLEM

Abstract. The customer contacts taking place after a sales transaction and the services involved are of increasing importance in contemporary business models. The responsiveness to service requests is a key dimension in service quality and therefore an important succes factor in this business domain. This responsiveness is of course highly dependent on the operational scheduling or dispatching decisions made in the often dynamic service settings. We consider the problem of optimizing responsiveness to service requests arriving in real time. We consider three models and formulations and present computational results on exact solution methods. The research is based on practical practical work done with the largest service organization in The Netherlands.

Date: December 15, 2008.

1. Introduction

The Traveling Salesman Problem (TSP) in which the salesman aims to find a minimum length route along customers he has to visit, is certainly among the most classical problems in operations research. Early work dates back to the 1930's and a first computational study is reported as early as 1954 [Dantzig et al.1954]. It has subsequently proven to be applicable to a wide range of different problems in manufacturing and service, and played and plays an important role in theoretical developments in combinatorial optimization and computer science. Very few of the applications address the domain where the problem was named [Garfinkel 1985], the efficient routing of salesmen. Moreover, the marketing paradigm it appears to fit is sales transaction oriented, and the objective is clearly efficiency oriented in its aim to reduce the travel time of or cost of the salesman. Hence the TSP tacitly assumes a product based marketing concept, called 'Old Marketing Concept' [Gummesson 1987] which has been replaced by long term relationship focused marketing concepts in which services play an dominant role. Nowadays the initial product sale is viewed as the start of a value creating relationship, not as a goal in itself. As Gronroos [1994] points out, production and service operations must therefore not be measured using performance indicators which are internally and efficiency oriented, thereby neglecting the impact on quality and customer satisfaction .The motivating application for the TSP is an illustrative example of such an internally and efficiency oriented approach.

As larger parts of the value created by companies, and indeed of our GDP are generated by the service industry, practitioners and researchers alike have worked to remedy this situation and develop new methods and concepts. Kaplan and Norton [1992] already stress the importance of putting customer oriented indicators in place and linking them with operational and financial indicators. For service operations, the objectives must therefore be to provide the quality needed to satisfy and retain customers. This holds true not only for pure service organizations but also for companies which are traditionally viewed as manufacturing companies. The professional printer/copier industry, where customers nowadays pay monthly for the services provided by a machine rather than purchasing it provides and illustrative example. Customer satisfaction is important to maintain the relationship, in which the customer asks for support when needed. Hence responsiveness to support requests becomes important, as is confirmed by the widely accepted model of Parasuraman et al. [1985] in which responsiveness is one of the five dimensions of service quality (next to reliability, assurance, empathy, and tangibles). Johnston [1995] considers a longer more explicit list of determinants of customer satisfaction, where the items are tested for their relationship with customer satisfaction and dissatisfaction. Responsiveness is again shown to be an important source of both satisfaction and

dissatisfaction. Although according to Parasuraman et al. [1985] customer satisfaction is by and large dependent on the perceived quality, rather than the delivered quality, Davis and Heineke [1998] show that in retail bank setting, actual waiting time is in fact a key driver of customer satisfaction. Likewise, Collier and Wilson [1997] show that failure to repair mobile phone connectivity within one day (responsiveness), or to arrive on time for the repair appointment (reliability) influence customer satisfaction and show how the relationship can be used to improve customer satisfaction by improving operations. Brady and Cronin [2001] and the references therein provide further evidence on the relationship between waiting time and customer satisfaction, and show that perceived waiting time is important to the provision of superior service quality.

Whether it regards after sales services which follow up on an initial sales transaction, or completely service based customer relations, responsiveness is to a large extent dependent on the operational decisions regarding the routing of service men. Routing service men to arrive as soon as possible in response to customer service request can therefore be viewed as a contemporary version of the TSP. In addition, it is closely related from a mathematical viewpoint and theoretically challenging. This paper discusses, models, and solves the resulting after salesman problem (ASAP) in a real time responsive context, as required in state of the art service operations. We compare models and solution methods by presenting computational results on test instances which have been constructed in collaboration with ANWB, the largest service organization in the Netherlands. Before doing so, the next section takes a more detailed view of the practical developments in this context, and reviews the literature on related models and solution methods.

2. Problem description and literature review

Essentially, this paper considers the following problem: a service providing company responds to service requests placed by its customers by sending a service man to the location of the customer request. We consider a general setting where customers might be mobile themselves, e.g. in their leased car, and hence the requests come from a wide variety of a priori unknown locations. The task of the service company is to respond quickly and satisfactorily to the service requests that pop up at these locations. Although the perception of responsiveness is typically subjective and dependent on circumstances [Davis & Heineke 1998], service contracts typically contain a service level agreement (SLA) specifying response times the service company has to fulfill. Typically, the SLA defines one or more of the following: a maximum waiting time until arrival of the service men, a time until repair in specific cases, specification of the service provided in specific cases, and/or an 'up time', which bounds the percentage of total time that the acquired product or service is not functioning properly. Since absolute waiting

time is important to customer satisfaction, the SLA can be designed to offer response times which fit the needs and expectations of the customers. In doing so, customer satisfaction will result if all requests can be assigned to service men such that the service is provided within the boundaries set by the SLA.

A service provider has several instruments at its disposal to realize the performance specified in the SLA. Of course, keeping a well trained, well equipped, and appropriately sized workforce of service men on the road is important. Secondly, it is important to process the customer calls adequately, and in case special services are required, it is important to understand the nature of the required service before sending a service man. In this paper however, we consider the crew size as fixed and assume that determining the nature of a service request is adequately dealt with. Instead, we focus on the following operational dispatching problem: given a set of service men and their starting locations, dynamically decide on an assignment of service men to service requests, which reveal themselves in real time, so that responsiveness satisfies the SLAs.

An important characteristic of this practical operational problem is of course that the input reveals itself in real time, and hence part of the input typically arrives after dispatching decisions have been taken (which establishes a major difference with the traditional TSP). In the literature, see e.g. [Psaraftis 1995], such problems are classified into classes such as *dynamic*, *on line*, or *real time* problems. In the problem under consideration, the following entities play a role:

(1) **Service requests** Over time, each service request consists of the following data:
   (a) Arrival time,
   (b) Location,
   (c) States *open*, *assigned*, *being served* and *completed*. The initial state after arrival is *open*. *Open* changes to *assigned* whenever a service man is assigned to service the request. It subsequently changes to *being served* when the service man arrives and starts servicing, and the state changes to *completed* when the service man is done.
   (d) Waiting time: the length of the time interval between the arrival time and the moment in time at which the status changes to *being served*, or if the state is still *open* or *assigned*, until the current time. Thus, for *open* and *assigned* service request it refers at any moment in time to the present waiting time. For other service request it refers to a registered waiting time.
   (e) Service time: the length of the time interval during which the service request is or has been in state *being served*. Thus, for service request in state being served it refers at any moment in

time to how long it is in state *being served* already. For completed service request it refers to the registered service time

(2) **Service men** Over time, each service men can be described by the following data:
   (a) Location,
   (b) Status: the status can be *available*, *traveling*, or *busy*. If a service man is in state *available*, his state can change to any of the other states immediately. If a service man is in the state *traveling*, he is traveling from an *origin* to a *destination*. The destination is usually the location of a service request. After traveling, a service man's state may change to *busy*, when his location coincides with the location of an *assigned* service request. When the state of the service man changes to *busy*, the state of the service request changes to *being served*.
   (c) Planning Status: Service man who are traveling or busy may have a subsequent service request already assigned to them in the planning. Hence the planning status variable may take on the values *plannedRequest* and *noPlannedRequest*. Whenever a service man with planning status *plannedRequest* completes servicing a request its status changes from busy to travelling. He starts traveling towards the request, taking the location of the just completed request as the origin and the location of the plannedrequest as the destination. His planning state changes to *noPlannedRequest*. The request changes status from *being served* to *completed.* If the planning status of the service man at the moment of completing the service is *NoPlannedRequest*, his status changes to *available* The status of the service changes from *being served* to *completed*.

Unless explicitly mentioned otherwise, we assume that the decision to assign a service request to a service man is irreversible. The decision to assign a service request to a service man is called an *assignment*. If an assignment decision is made regarding a service man in state *available*, and who therefore has planning state *noPlannedRequest*, his state will change to *traveling*. He will start traveling with his current location as origin, and the location of the assigned service request as destination. If his state is *traveling* or *busy*, and his planning state is *noPlannedRequest*, the assignment will cause his planning state to change to *plannedRequest*. Service men whose planning state is *plannedRequest* cannot be assigned. The length of the service time, or remaining service time, may not be a priori known. Hence it is specified by a service time distribution function. Notice that, while a service request is *being served*, its remaining service time can be conditional on the present service time. Throughout this paper we assume that the travel

time for a service man can be computed as a function of the corresponding origin-destination pair.

Throughout this paper we will refer to an *on line* instance as being defined at a certain *time instant* as follows:

(1) service request that have arrived at or before that time instant
(2) service men
(3) a distance function (or matrix) providing the time required to travel between relevant (origin,destination) pairs.
(4) For each open service request, a distribution of expected service time duration
(5) For each service request being served, a distribution of expected remaining service time duration
(6) All assignments of service requests to service men already made.

We define $t = 0$ to be the first time instant, and $t = T$ to be the time instant at which the last service request arrives. The *off line instance* is an extension of the on line instance at time 0. It consists of all data of the on line instance at time 0, plus the specification of all service requests arriving until T, and the durations of the service times. Hence in the off line instance, all problem data are known in advance. In this paper by contrast, we focus on approaches to solve the ASAP by repeatedly solving models of on line instances to optimality, and discuss the quality of the thus obtained solutions for the off line instance. The solution which is thus obtained for the off line instances is called the *end of day* solution. The quality of the solution will be measured in registered waiting times, service times of the service requests, and travel times of the repair men. Notice that the optimal solution of the off line instance provides a lower bound value of the end of day solutions obtained through repeatedly solving on line instances. The worst case ratio between the two is known as the competitive ratio, or the price of information (see for instance Sitters et al. [2003] for a related reference in multi server scheduling). A final problem which is worth defining is the stochastic off line problem. It is the variation of the off line problem where the arrivals of all service requests are known in advance.

Psaraftis [1995] and Gendreau and Potvin [1998] give a clear and thorough overview of work on dynamic traveling salesman problems. Psaraftis explicitly points out that as technology proceeds, the importance of dynamic versions can be expected to increase. Indeed, as technology to locate service men, customers and service requests in real time has become easily accessible, improving responsiveness has become a competitive weapon in various service industries. Psaraftis also points out that in such settings the traditional objective function of the TSP, to minimize total travel time, is not very satisfactory, and proposes to minimize mean system time (average waiting time). We will consider this issue in more detail below.

An important line of related research addresses multiple server problems (see [Bertsimas & van Ryzin 1991] for a seminal reference). In such server

problems, the goal is to assign service requests to servers, often with the objective to minimize mean system time. In server problems the requests arrive dynamically, and the assignment decisions are referred to as policies, or dispatching rules. Simple policies decide which server to assign to a request upon arrival of the request, but more advanced policies in which the assignment is delayed, even if a server is available, are common. Under these delaying policies it is possible to group service requests and assign them in a specific order to servers. Bertsimas and Van Ryzin [1991] analyze stochastic and dynamic vehicle routing problems which are closely related to the problem proposed in this paper from a multi server perspective. They prove the existence of policies which are guaranteed to deliver a finite, or bounded average waiting time, whenever possible. For the single server problem simulation results [Bertsimas & Van Ryzin 1991b] indicate that a simple Nearest Neighbor policy outperforms other policies such as space filling curve based policies, in which assignments are delayed and service requests grouped. Lu et al. [2002] prove that for specific heavy traffic problems an approach which partitions the region into subregions per server in which a TSP based delay policy is utilized, is optimal. Irani et al. [2001] provide competitive ratio analysis for the same problem with the objective of maximizing the number of requests serviced by their deadline (as it may be specified in the SLA).

There is an active more practice oriented line of research on repair men scheduling problems. Much of the research on practical problems (see [Gendreau et al. 1998] for references) uses local search and metaheuristics. These approaches differ fundamentally from the server based policies in the following way. The assignment decisions made for subsequent on line instances can be constructed by making flexible use of assignment decisions for the previous on line instance, while taking the newly arrived service requests into account. Techniques used are insertion policies, or reptimization which takes the previous solution as a starting solution. Moreover, several authors, see e.g. [Ichoua et al. 2000] consider the possibility of diversion, which entails changing the destination of an already traveling service man, thereby undoing assignments.

Many of the practical applications are from a context which differs from the real time service response delivery considered in this paper. Most of the applications stem from logistic contexts, where the requests regard a pick up or a delivery of goods. Such partially dynamic problems are for instance discuss in [Ichoua et al. 2006]. Naturally, in such problems customers have different expectations; service is often not requested as soon as possible, but within a certain time window which opens in the future. The different nature of the problem also leads to different operational priorities and objective functions.

Krumke et al. [2001] have considerably advanced practical work by taking an approach which is based on solving on line instances to (near) optimality using a set partitioning formulation that is solved by column generation and

branch and bound. The objective function they consider consists of three parts. Firstly they incur costs per kilometer traveled by a service man. Second, they set a bound on waiting time, and incur a penalty costs that is linear in the waiting which exceeds the bound. This models captures agreements as in the aforementioned SLAs. Thirdly, they consider using extra servers, subcontractors, to whom requests can be assigned as well, but at a certain cost per request. Krumke et al. [2001] report extensive computational results on large real life on line instances from the German automobile club ADAC, which can be quickly solved to (near) optimality with the column generation approach. Further they report that this approach outperforms various metaheuristics, and that the relative performance gets better when the arrival rate of service requests increases Added to our emphasis on modeling, this has lead us to the decision to disregard metaheuristics and local search approaches in the analysis.

In view of the importance of responsiveness, it is remarkable to notice that only very few authors (see for instance [Gendreau et al. 1998], [Irani et al. 2001], [Krumke et al. 2001]) consider objective functions which are customer oriented. Most of the literature considers the average waiting time. Minimization of the average waiting time and/or the cost are of course reasonable objectives, but clearly refer to internal operational performance indicators. They are not relevant to individual customers whose expectations and satisfaction relate to meeting their expectations, and indeed to their own waiting time. Recent customer satisfaction research by the Dutch Automobile Association [ANWB 2003] shows that dissatisfaction typically stems from waiting excessively long or from waiting longer than promised and reveals no indications that it is otherwise linked to average waiting time. Hence, there is a need to model and solve service men dispatching problems with a customer satisfaction objective, as this paper strives to achieve.

The outline of this paper is now as follows. In Section 3 we first enlist and compare three natural approaches to solve the end of day problem by repeatedly solving models for on line instances to optimality. Section 4 compares and evaluates basic version of the thee approaches by comparing end of day objective values It turns out that the set partitioning approach dominates the other approaches. In Section 5 we consider the sensitivity towards the reoptimization frequency. Section 6 addresses the choice of objective for the models to solve the on line problems. Section 7 studies the value of allowing diversion, the rerouting of already assigned service men. Section 8 explores the value of having perfect information on the service durations. Section 9 discusses the results, and formulates an agenda for further research on ASAP.

3. Modelling and Solution Approaches

We start our modelling efforts by mathematically capturing the objective of customer satisfaction for so far as it relates to dispatching. As explained,

customer satisfaction is directly dependent on the perceived and actual responsiveness. In the setting under consideration responsive service regards the timely arrival of a service man, and subsequently adequate service so that the total discontinuance of the customer is as short as possible. Natural performance indicators to be included in the SLA are therefore waiting time until arrival of the service man and sum of waiting time and service time. The latter is especially relevant when the service man is needed to repair a device that is not delivering the service that the customer is entitled to. In many occasions however, the quality of the activities of the service man are hard to judge for a customer, and their duration is often accepted as given, provided that the service man makes a professional impression and remedies the problem [ANWB 2003] - relating to the assurance dimension of service quality [Parasuraman et al. 1985].Since we focus in this paper on the dispatching decisions, we will not analyze or address the duration of the activities by the service man explicitly. Instead we consider it to be a given stochastic variable. Focusing subsequently on the responsiveness as it results from dispatching decisions, the waiting time until arrival of the service man forms the predominant driver of customer satisfaction. In our basic models for capturing customer satisfaction we therefore assume the SLA to specify a *waiting time threshold* regarding the maximum allowable waiting time until arrival of a service man. We return to the relevance of the sum of waiting and service time in subsequent sections.

The following list of performance indicators is a mix of objectives encountered in scientific literature and practice:

(1) Total distance traveled by all service men,
(2) Total idle time over all service men, where idle time is the time that a service man is in state available,
(3) Average number of incidents per hour per service man,
(4) Sum of the completion times of the requests,
(5) Average waiting time per request,
(6) Maximum waiting time of a request,
(7) Number (or percentage) of customers whose waiting time is within the SLA threshold,
(8) Sum over all customers of the waiting time in excess of the SLA threshold,
(9) Weighted sum over all customers of the waiting time in excess of the SLA threshold, where the weight is a nondecreasing function of the excess waiting time.

Clearly, the first of these indicators are completely process oriented, subsequent ones are more market oriented, and the SLA related ones become more customer oriented. We will see in subsequent sections that in order to minimize an end of day objective function for an approach which repeatedly solves on line problems, it can be advantageous to select a different objective function for those on line instances. Hence, despite the fact that

some of the objectives might conflict with customer satisfaction, all of the aforementioned objective functions are kept under consideration.

Let us briefly consider instances in which the arrival rates of requests is high enough to avoid idle time of the service men. Taking into account that the service durations are given, the zero idle time implies that the problem boils down to managing the travel times of the service men. It is not hard to see that in the absence of idle time, a solution which minimizes the sum of the travel times of the service men, minimizes the sum of the waiting times of the customers. Hence it minimizes the average waiting time, and since service durations are exogenous, it minimizes the sum of the completion times. Thus for heavy traffic instances, which we will argue to be more important to practical settings, in which idle time is unlikely, objectives 1, 3, 4 and 5 are very closely related.

While discussing the objective functions, let us also mention that at the time of making an assignment decision the resulting waiting times are not necessarily exactly known. For instance the assignment might involve a service man who is travelling to its current request, still has to service the request which has a stochastic duration, and then to travel to the next assigned request. Obviously, his arrival time at this next request depends on the service duration of the current request. Hence, the assignment decisions are typically based on expected arrival times, which are calculated as follows:

(1) If the service man is available, the expected arrival time equals the travel time when taking his current position as the origin and the position of the request as the destination,
(2) If the service man is busy, the expected arrival time equals the expected remaining service time at his current position plus the travel time when taking his current position as the origin and the position of the request as the destination,
(3) If the service man is traveling, the expected arrival time equals the remaining travel time to his current assigned request, plus the expected service time of the current assigned request, plus the travel time when taking the position of the current assigned request as the origin and the position of the next (planned) assigned request as the destination.

We will consider various objective functions, some of which are multicriteria functions. If the assignment objective function involves waiting times, the expected waiting times can be deduced from the above defined expected arrival times.

The expected remaining service time can of course only be computed when a distribution function is assumed on the service time. In our experiments, we use an exponential service time distribution function. This distribution captures the real life data well, yet has the counterintuitive property that the expected remaining time doesn't decrease as service is in progress.

We consider three solution approaches for real time service men scheduling. The basic idea of all three approaches is to solve the problem by solving a series of on line instances exactly yet by different models. We compare the performances by comparing the end of day objective values. We use the following models:

(1) **Multi server FCFS:** Process the new requests periodically at a First Come First Served (FCFS) basis. Although the simple FCFS may deliver non optimal solutions, it has the property that customers are processed in the order in which they arrive, which results in a fairness that is often highly appreciated by service customers. Moreover, this approach serves as a simple reference case against which the other approaches will be benchmarked. Let it be noted however that it is known to perform poorly in a single server setting [Bertsimas & van Ryzin 1991] when compared to more advanced server policies, such as space filling curves, or nearest neighbor.
(2) **Matching:** Periodically reconsider simultaneously the assignments of requests planned as next requests for the service men, and make a new assignment of such planned requests for service men. In principle current requests to which service men are travelling or that they are servicing are not reconsidered. Instead the reoptimization only concerns planned requests, with the restriction of planning at most one request per service man. This resulting optimization problem can be modelled using a bipartite graph where the service men form one color class, and the request the other. Finding an optimal solution can subsequently be defined as a linear assignment problem.
(3) **Set partitioning:** Periodically reconsider the assignments of planned requests, without the restriction that there is at most one planned requests per service men. Instead, we require that all requests are assigned. This service men may have a number of requests assigned to them, in which case an order in which the requests are processed needs to be determined to be able to compute an objective function value. This model requires to partition the set of all requests into subsets and to assign the subsets to the service men, so that each service man is assigned at most one subset. To calculate the waiting times, a routing problem of the requests per service man must be solved. Thus we establish the formulation of the problem as a set partitioning problem.

The FCFS implementation operates as follows. Periodically, the newly arrived requests are processed in order of appearance. While there is at least one open request and at least one service man with planning state NoPlannedRequest, we repeatedly assign to oldest open request the service man which yields lowest objective value (waiting time, travel time, et cetera). When an assignment is made, the status of the request changes to assigned.

If the service man is traveling or busy, its plannedRequest state changes to PlannedRequest. If the service man is available, its state changes to traveling. Notice that this policy doesn't only choose among idle service men. On the other hand it doesn't assign more than one future service requests to each of the service men. It is easily checked that the running time to solve an on line instance is linear in the number of open requests times the number of service men.

The second solution approach we consider is based on a bipartite graph formulation for the on line problem. It has been successfully applied for several years by ANWB. Moreover, it a natural generalization of the Nearest Neighbor policy which is reported to perform well in the single server problem [Bertsimas & Van Ryzin 1991b]. For a given time instant, a bipartite graph $G(V_1, V_2, A)$ is generated as follows. Vertex class V_1 contains a vertex for every service man whose planning state has value NoPlannedRequest, and vertex class V_2 contains a vertex for each service request with state open. The bipartite graph is complete, and the cost of arc $(v_1, v_2, v_1 \in V_1, v_2 \in V_2)$ can be defined using any of the objective aforementioned objective functions (waiting time, travel time, et cetera), or a combination of them. Thus we establish the formulation of the on line problem as a linear assignment problem. The Hungarian method solves it in time polynomial in $(\min\{V_1, V_2\})^3$ time, and it 's time complexity is therefore cubic in the number of request and service men. This matching approach aims to make dispatching decisions in real time by finding a simultaneously assigning work to all service men with planning status NoPlannedRequest. Casting the FCFS approach in this context, leads to its interpretation as a greedy heuristic to solve the linear assignment problem which the matching approach solves to optimality. Other server policies can be interpreted likewise. Let it be noted however that repeatedly solving the assignment problem to optimality doesn 't guarantee optimal end of day performance. In fact, finding exact solutions to a model used for solving on line instances doesn't necessarily lead to better end of day solutions than solving it heuristically. Thus it remains to be seen that the matching approach yields better solutions than the FCFS approach.

A similar argument holds with respect to the set partitioning approach discussed below. It is more complete in its modelling of the on line instances, but that doesn't necessarily imply that it provides better end of day performance. The optimality of the solution for linear assignment model for an on line instance has its appeal, but the model has shortcomings as well. Even if the on line instance solved is the instance at time T, in which case no further input data arrives, it doesn't necessarily provide an optimal solution. An important characteristic of the approach is that it never assigns more than one service request to a service men which has planning status noPlannedRequest. Actually it can be viewed to work from the perspective of the service men, a resource perspective, rather than from the perspective of the customers. In particular this implies that when there are more open

requests than service men with NoPlannedRequest, the model solution disregards waiting times of customers who are not in the optimal assignment. Since this situation is likely to occur under busy circumstances, this property might easily turn out to be a shortcoming. We now present the set partitioning approach developed by Krumke et al. [2001], which does take a customer perspective. It maintains a planning in which all customers are assigned to service men, perhaps more than one customer per service man. Moreover, the solution it provides for the on line instance at T, is the optimal solution for the equivalent stochastic off line instance.

Rather than making pairs of service requests and service men, Krumke et al. [2001] solve the on line instances by constructing a solution in which the set of service requests is partitioned into subsets, and in which the subsets are assigned to service men. Thus, every requests is now taken into account in the solution. The value of a solution depends of course on the order in which each of the service men services the requests in the subset assigned to him, and the objective function chosen. Krumke et al. [2001] show how these combined partitioning and routing problem can be solved using integer linear programming techniques. They consider a set partitioning model, which uses the set O of all open service requests, and the set M of all service men. Hence $M \cup O$ is the set of all service requests and service men, and this is the ground set of a set partitioning formulation. The subsets by which the ground set is to be partitioned are the subsets of $M \cup O$ in which there is exactly one service man and zero or more service requests. Let J be the set of these subsets, and let $a_{ij} = 1$ if for $i \in, M \cup O$, $j \in J$, request or service men i is contained in subset j. The well known integer linear programming formulation for the set partitioning problem then reads:

$$\min \sum_{j \in J} c_j x_j \tag{1}$$

$$\text{s.t.} \tag{2}$$

$$\sum_{j \in J} a_{ij} x_j = 1 \forall i \in O \tag{3}$$

$$x_j \in \{0, 1 \forall j \in J\} \tag{4}$$

We denote this problem by SP and its value is denoted $v(SP)$. Determining the cost of each column j is a nontrivial task. let S be the set of service requests covered by column j. Then, the current location of the service man, the current waiting times, and the locations of the service requests in s are required to calculate the cost c_j. Subsequently, c_j is defined to be the minimum cost over all routes starting from the current position of the service man, along the service requests in s. The cost of a route can be determined as before, using expected arrival times. For instance, if the cost function is to minimize the expected maximum completion time, calculating the cost of a route requires to solve a classical TSP. Of course the TSP is NP-Hard

to solve and so is the calculation of the costs on all objective functions we consider. Since set partitioning problems in general are hard to solve as well, and the number of subsets is exponential in the input size, the formulation given above is not easily formulated and solved. However, Van de Klundert et al. [2008] and Krumke et al. [2001] report satisfactory performance in real time settings for large scale problems, based on the following solution approach proposed by Krumke et al. [2001]. More specifically they show that on line instances arising in the planning of nationwide operating service organizations in Germany and The Netherlands which service millions of customers per year can be solved in a matter of seconds.

Let X be any of the aforementioned subsets of $M \cup O$, and let $v(SP|X)$ be the value of the restricted version of SP in which only the columns in SP are feasible. Then obviously $v(SP|X) \geq v(SP)$. Moreover, let $R(SP)$ and $R(SP|X)$ be the linear relaxations of SP and $SP|X$ respectively. Then, it must hold that $v(R(SP|X)) \geq v(R(SP))$. Since the number of rows of $R(SP)$ is polynomially bounded, there exists an optimal solution consisting of a polynomially bounded number of columns (which can even be found in polynomial time [Grotschel et al. 1980]). A well known technique to find the optimal solution to $R(SP)$ quickly is column generation, in which one starts with some feasible solution and subsequently adds columns until a solution X* is obtained for which strong duality holds and hence $v(R((SP|X\hat{\ }*)) = v(R(SP))$. Once this solution for $R(SP|X\hat{\ }*)$ is constructed, a feasible solution to 1 is constructed by finding the optimal solution for $(SP|X)$ by branch and bound. In terms of the service men routing problem under consideration, there are many implementational issues to consider, especially in the construction of tours of requests with negative reduced costs, for which we refer to [Krumke et al. 2001], and [Wormer 2005].

Let it be noted that it may hold that $v(SP|X) > v(SP)$, and hence that the thus constructed solution is not optimal for 1. Indeed, the condition that $v(R((SP|X\hat{\ }*)) = v(R(SP))$ is necessary but not sufficient for $v(SP|X\hat{\ }*) = v(SP)$. Thus, in the end, the approach solves the Set Partitioning formulation of the on line instances heuristically. Nonetheless, [Krumke et al. 2001], and [Wormer 2005] report very small gaps between the solution thus obtained and the optimal solution, and in fact also small integrality gaps. Knowing that solving the ASAP by solving a series of on line instances will not provide optimal solutions anyway, we accept this near optimal solution as the outcome of the set partitioning model.

4. Basic computational results

A first analysis is simply to review the three different solution approaches in a fairly basic setting. To this purpose, we developed a simulation model in which each of the solution approaches can be tested on identical data sets, and we implemented each of the three approaches. Our basic simulation set up is as follows. The service area will be a square of 125 by

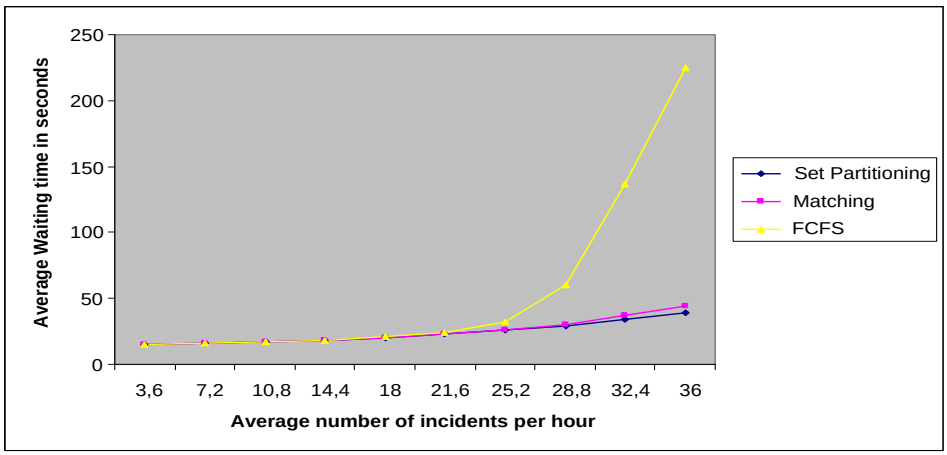

FIGURE 1. Base results. model comparison

125 kilometers, in which 20 service men operate who travel at a speed of 1 kilometer per minute. The travel distance between any two points in the service area is the euclidean distance. In the base scenario, the planning is reoptimized every 30 seconds. The service time duration is exponentially distributed with an expected service duration of 15 minutes. These data resemble the rural areas of the practical application which motivated the research ([Huigenbosch et al. 2008]).We have generated 10 instance types, which vary according to the probability of new service requests arriving in the next 10 seconds. This probabilities will range from 0.01 to 0.1 in equally sized steps. The expected number of requests per hour is therefore between 3.6 and 36. For every set of parameters we have generated 20 instances of 24 hours, or 20 days. The presented results are the averages over the 20 days per parameter set.

The base results presented in figures 1-3 regard average waiting times minimization: the on line instances are solved using an average waiting time objective in the FCFS, matching and set partitioning approaches. For the matching model this means that the objective is to find a maximum cardinality matching which minimizes the sum of the corresponding waiting times, and for the set partitioning model this means that the costs of a tour equals the sum of the waiting times of the customers in a tour. Figure 1 displays the resulting end of day average waiting times. We don't propose that average waiting time is the most important performance indicator, but use it here to provide a first comparison between the three approaches. Figure 1 clearly displays that the FCFS dispatching rule will result in long average waiting times in instances where the other two models provide much better results. Figure1 also indicates that the matching model and the set partitioning model provide comparable solutions in terms of average waiting time,

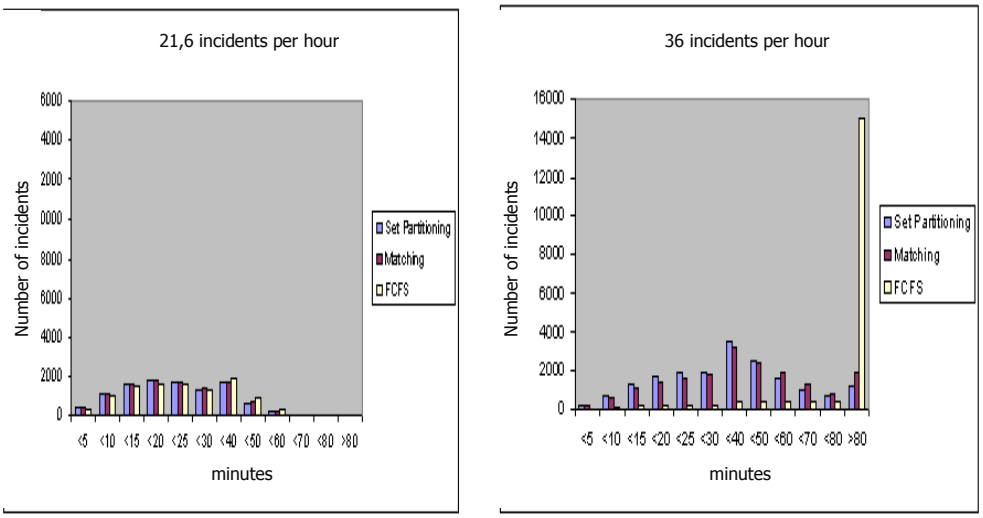

FIGURE 2. Base results,waiting time distributions

but that the set partitioning approach performs better when the request intensity increases.(Such instances with high arrival rates will be called heavy traffic [Bertsimas & van Ryzin 1991] instances in the remainder). More specifically, the average waiting times are identical for request arrival rates from 3.6 to 20.6, after which the average waiting times difference increases as the arrival rate increases. For an arrival rate of 36, the average waiting times are 38.8 minutes for the set partitioning approach and 43.9 for the matching approach, a difference of more than 10 percent. This difference might be explained by the fact that when request arrival rate increases, the matching model is more likely to have a number of open requests that exceeds the number of repair men with status noPlannedRequest, and therefore leaves requests unassigned in its optimal solution. The set partitioning model in which all requests are always assigned to service men explicitly addresses the difficulties of these scenarios. Figure 2 reveals that the end of day performance differences are even bigger when considering the waiting time in excess of a waiting time thresholds of e.g. 30 or 60 minutes. For a request arrival rate of 36 per hour, the FCFS has 87.7 percent of customers waiting more than 60 minutes. The matching approach has 22 percent of customers waiting more than 60 minutes, while the set partitioning approach has 15.3 percent of customers waiting more than 60 minutes.

Figure 3 provides further insight into the operational performance drivers of the responsiveness displayed in Figures 1 and 2. The FCFS approach tends to assign requests to service men which are relatively far away as the intensity increases, leading to an decrease in productivity while demand requires an increase. In the other models the driving time also tends to

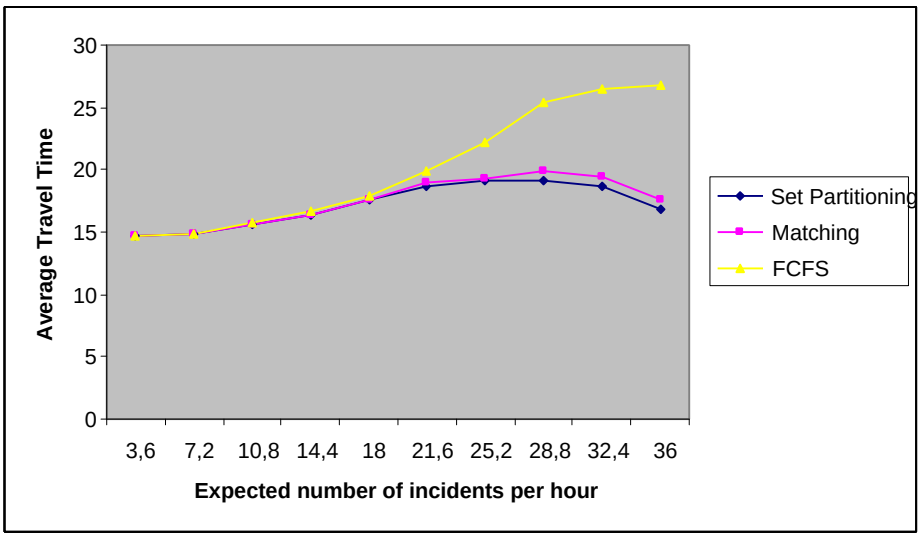

FIGURE 3. Base results, travel time comparison

increase, because of the decreased likelihood of a nearby idle service man as the arrival rate increases. When it increases further, the next open request for a service men that has no planned requests is more probable to be close by, and hence the average travel time decreases (albeit as we concluded from Figure 1 the average waiting time continues to increase). As the arrival rate increases to 36, the average travel time of the FCFS approach continues to increase, whereas the other two approaches enjoy a decrease. We notice again that the set partitioning approach outperforms the matching approach under these circumstances, but with a smaller relative difference (of less than 4 percent).

The results with relatively higher arrival rates are in our view more important. When the arrival rate is low, many requests are serviced by a nearby service man with state available. Any reasonable method will provide a good solution in this case. Customer satisfaction is not at stake, but operating costs per served request might be fairly high. Typically, price based competition forces companies to reduce service operating costs to the minimum level which results in the aspired customer service level, entailing that scenarios with more challenging request arrival rates are more relevant than the scenarios which are easier to handle. It is exactly under these circumstances when the added value of appropriate dispatching methods pays off by enabling more competitive combinations of customer service levels and operating costs. Hence its inability to successfully solve such instances disqualifies the FCFS approach. The fairness of servicing all requests in order of arrival eventually leads to longer expected waiting for all requests.

Thus we consider the FCFS approach inferior to the matching and set partitioning approach. Moreover, we conclude that the latter performance better especially when considering waiting in excess of the SLA threshold value for heavy traffic instances. Since the solution approach for the set partitioning formulation of the on line instances allows to solve them almost to near optimality within seconds for large scale real life instances, the remainder focuses on the set partitioning approach.

5. Reoptimization frequency

An issue which is worthy of attention in designing solution approaches for real time service men scheduling problems, is the frequency of (re)solving the on line instances. A natural moment to consider reoptimization is whenever new input data arrives, e.g. whenever a new request arrives. In addition, reoptimization might be considered when new service men become available, or when service requests are completed. Despite various authors [Krumke et al. 2001][Huigenbosch et al. 2008] reporting to successfully solve large real life on line instances within seconds, the required reoptimization frequency might be too high. Another reason to consider a lower reoptimization frequency is that as reoptimization is postponed, more input data arrives, allowing for a more informed, better solution. Of course, there is a limit to benefitting from such postponement since the waiting time of the newly arrived requests increases while they are not taken into consideration. Krumke et al. [2002] report that in the practical instances they considered little difference in outcomes was observed when decreasing the reoptimization frequency to once per minute. Figures 4,5 and 6 display results for the set partitioning approach. We observe that reoptimizing every 10 seconds - although perhaps unrealistic - improves average waiting slightly. It has a higher percentage of customers in fast response categories of less than 5 and less than 10 minutes, but performs comparable regarding long waiting. Waiting 60 seconds before reoptimization results in decreasing the number of customers in fast response classes, as clearly shown in Figure 7, but reduces the number of customers waiting long, and therefore improves end of day SLA performance. The average waiting times over all 30 and 60 second reoptimization scenarios are equal. Hence we conclude that for heavy traffic instances an extra delay of the dispatching decisions by 30 seconds is more than made up for in operational performance.

We conclude that small improvements can be obtained by reoptimizing quite frequently under light traffic conditions, and less frequently under heavy traffic conditions. In the remainder we nevertheless use the fixed reoptimization interval of 30 seconds as used in the base scenarios of the previous section.

Set Partitioning 10

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	8,0	22,8	24,3	20,7	13,0	6,2	3,7	0,8	0,2	0,1	0,1	0,0	15	95,0	99,7
7,2	8,7	19,2	26,1	19,8	12,9	7,3	4,5	1,1	0,2	0,1	0,0	0,0	15,5	94,0	99,8
10,8	6,7	18,2	23,0	21,7	14,2	8,1	6,1	1,5	0,3	0,1	0,0	0,0	16,6	91,9	99,8
14,4	5,9	16,1	21,2	20,3	16,1	9,5	7,8	2,1	0,7	0,1	0,1	0,1	18	89,1	99,7
18,0	4,9	13,9	18,4	18,6	15,4	11,7	11,4	3,9	1,0	0,5	0,2	0,2	20,1	82,9	99,2
21,6	4,0	10,5	15,6	16,6	16,1	12,4	15,1	6,2	1,9	0,8	0,3	0,3	22,8	75,2	98,4
25,2	3,1	9,1	12,6	15,2	15,0	13,3	17,1	8,4	3,5	1,5	0,7	0,6	25,6	68,3	97,3
28,8	2,2	6,6	10,5	12,7	13,9	13,2	20,2	10,7	5,0	2,5	1,2	1,3	29,1	59,1	95,0
32,4	1,3	4,6	7,9	10,5	12,2	12,1	20,5	13,7	7,8	4,4	2,3	2,8	34	48,6	90,6
36,0	1,1	4,0	7,1	9,2	10,5	10,8	19,3	14,1	8,7	5,9	3,4	5,9	38,4	42,7	84,8

Set Partitioning 30 (base)

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,3	22,9	24,7	20,3	13,5	6,2	3,8	0,8	0,2	0,0	0,2	0,0	15,2	94,9	99,7
7,2	8,1	19,4	25,6	20,0	13,1	7,5	4,8	1,2	0,2	0,1	0,0	0,0	15,7	93,7	99,9
10,8	6,1	17,9	23,1	21,4	14,9	8,2	6,4	1,5	0,3	0,2	0,0	0,0	16,9	91,6	99,8
14,4	5,6	15,5	21,1	20,5	16,3	9,4	8,1	2,2	0,8	0,2	0,1	0,0	18,3	88,4	99,5
18,0	4,7	13,7	18,1	18,7	15,7	11,8	11,4	4,0	1,0	0,5	0,1	0,0	20,2	82,7	99,1
21,6	3,9	10,3	15,5	16,4	15,9	12,2	15,7	6,2	2,3	0,8	0,5	0,3	23,1	74,2	98,4
25,2	2,8	8,8	13,0	14,7	14,3	13,5	17,3	9,1	3,6	1,6	0,7	0,7	26	67,1	97,1
28,8	2,2	6,6	10,6	12,8	14,1	13,4	19,8	10,4	5,1	2,6	1,2	1,3	29	59,7	95,0
32,4	1,3	4,3	7,8	10,6	12,1	12,4	20,6	13,4	7,6	4,5	2,2	3,2	34,3	48,5	90,1
36,0	1,0	3,9	6,9	9,1	10,7	10,7	19,3	14,0	8,9	5,4	3,6	6,3	38,8	42,3	84,5

Set Partitioning 60

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	6,5	22,0	25,5	20,2	13,9	6,5	3,9	0,9	0,2	0,0	0,1	0,1	15,4	94,6	99,6
7,2	7,3	19,3	25,3	20,7	13,3	7,5	5,0	1,1	0,3	0,1	0,1	0,0	16	93,4	99,8
10,8	5,5	17,6	22,9	22,0	14,9	8,6	6,5	1,4	0,3	0,2	0,0	0,0	17	91,5	99,7
14,4	5,1	15,1	21,1	20,2	16,2	9,8	8,8	2,5	0,8	0,2	0,1	0,0	18,6	87,5	99,6
18,0	4,4	13,1	18,2	18,8	15,6	12,0	11,9	4,0	1,2	0,5	0,2	0,1	20,5	82,1	99,2
21,6	3,4	9,9	14,9	16,7	16,3	12,5	16,1	6,3	2,5	0,8	0,5	0,3	23,5	73,7	98,6
25,2	2,4	8,7	12,2	15,2	14,3	13,3	17,9	9,0	3,9	1,5	0,8	0,7	26,3	66,1	96,9
28,8	1,9	6,4	10,5	12,7	14,2	12,7	20,3	10,5	5,5	2,6	1,3	1,3	29,4	58,4	94,7
32,4	1,1	4,5	8,0	10,6	12,0	12,3	20,7	13,6	7,4	4,2	2,3	3,3	34,4	48,5	90,2
36,0	1,0	4,1	7,5	9,7	11,2	11,2	19,8	14,1	8,6	5,3	3,0	4,5	36,8	44,7	87,2

FIGURE 4. Comparing different reoptimization periods for the set partitioning algorithm

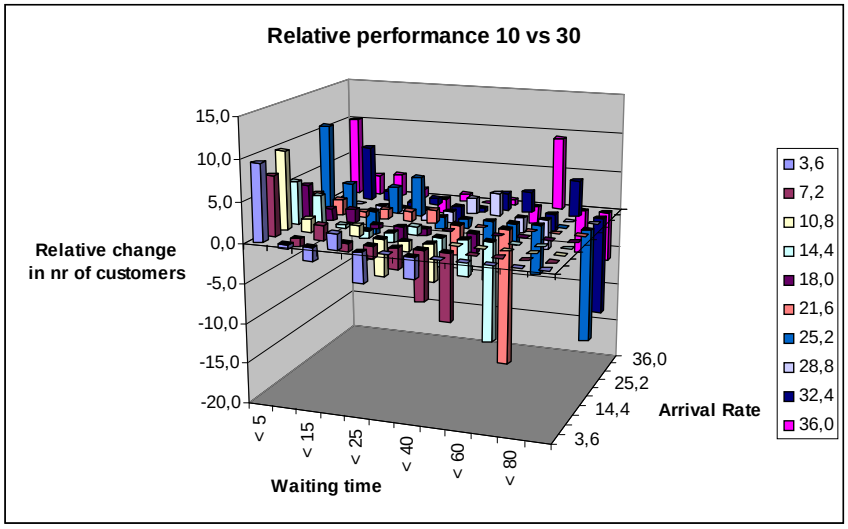

FIGURE 5. Relative performances for reopt interval of 10s to 30s

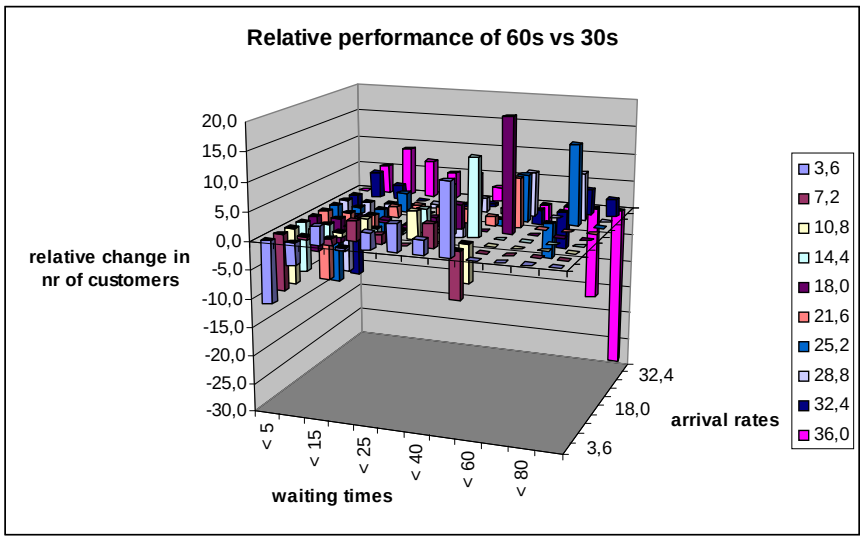

FIGURE 6. Relative performance for reopt interval of 60s to 30s

6. FINE TUNING THE ON LINE DECISIONS FOR END OF DAY PERFORMANCE

As will be demonstrated shortly, it is not true that minimizing a certain end of day objective, such as sum of the waiting times, is best achieved by using the same objective in the model by which the on line instances are solved. By consequence, finding the objective, or combination of objectives which are used in finding solutions for the on line instances to optimize an end of day objective is a non trivial task. We have not explored this issue from a theoretical viewpoint, as can for instance be done using competitive analysis, but report on our extensive computational experiments in this section. This section forms the core of the paper since it addresses the pivotal issue of steering operations for customer satisfaction. We have argued that customer satisfaction is best modeled by using SLAs regarding a maximum response time, and hence our aim is now to find objectives for the on line instances which result in minimum possible waiting in excess of the agreed service levels.

A first and insightful comparison is depicted in Figure 7 where we compare end of day waiting time distributions for four different objectives for on line instances as they are solved using the set partitioning approach.

The first group of results regards the base results already displayed in Figure 7. The objective for the on line instances is to minimize the total (or, equivalently, the average) waiting time and disregards SLA thresholds. The next two groups of results regard scenarios where the threshold plays an explicit role in the solution of the on line instances. Based on the practical work described in [Huigenbosch et al. 2008] we consider a SLA threshold

Waiting

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,3	22,9	24,7	20,3	13,5	6,2	3,8	0,8	0,2	0,0	0,2	0,0	15,2	94,9	99,7
7,2	8,1	19,4	25,6	20,0	13,1	7,5	4,8	1,2	0,2	0,1	0,0	0,0	15,7	93,7	99,9
10,8	6,1	17,9	23,1	21,4	14,9	8,2	6,4	1,5	0,3	0,2	0,0	0,0	16,9	91,6	99,8
14,4	5,6	15,5	21,1	20,5	16,3	9,4	8,1	2,2	0,8	0,2	0,1	0,0	18,3	88,4	99,5
18,0	4,7	13,7	18,1	18,7	15,7	11,8	11,4	4,0	1,0	0,5	0,1	0,0	20,2	82,7	99,1
21,6	3,9	10,3	15,5	16,4	15,9	12,2	15,7	6,2	2,3	0,8	0,5	0,3	23,1	74,2	98,4
25,2	2,8	8,8	13,0	14,7	14,3	13,5	17,3	9,1	3,6	1,6	0,7	0,7	26	67,1	97,1
28,8	2,2	6,6	10,6	12,8	14,1	13,4	19,8	10,4	5,1	2,6	1,2	1,3	29	59,7	95,0
32,4	1,3	4,3	7,8	10,6	12,1	12,4	20,6	13,4	7,6	4,5	2,2	3,2	34,3	48,5	90,1
36,0	1,0	3,9	6,9	9,1	10,7	10,7	19,3	14,0	8,9	5,4	3,6	6,3	38,8	42,3	84,5

Waiting + SLA

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,3	22,9	24,7	20,2	13,5	6,3	3,8	0,8	0,2	0,1	0,1	0,0	15,2	94,9	99,7
7,2	8,1	19,3	25,9	20,0	13,0	7,4	4,7	1,1	0,2	0,1	0,0	0,0	15,7	93,7	99,7
10,8	6,0	17,8	23,2	21,6	14,9	8,3	6,3	1,4	0,3	0,1	0,0	0,0	16,8	91,8	99,8
14,4	5,7	15,3	20,9	20,2	16,4	9,7	8,5	2,3	0,7	0,2	0,1	0,0	18,4	88,2	99,7
18,0	4,6	13,7	18,2	18,3	15,8	12,0	12,1	3,7	1,0	0,4	0,2	0,0	20,2	82,6	99,4
21,6	3,7	10,0	14,8	16,3	16,2	12,5	16,6	6,3	2,1	0,8	0,3	0,3	23,3	73,5	98,5
25,2	2,8	8,4	11,6	14,0	14,4	13,9	19,4	9,1	3,6	1,5	0,8	0,5	26,4	65,1	97,2
28,8	1,9	5,6	9,2	11,8	13,2	13,7	22,4	11,7	5,6	2,6	1,2	1,1	30,1	55,4	95,1
32,4	1,0	3,4	5,9	9,4	12,0	13,2	23,6	14,8	7,9	4,0	2,3	2,6	34,9	44,9	91,2
36,0	0,6	2,5	4,6	6,6	9,4	11,8	23,0	16,2	9,8	5,9	3,9	5,6	40,3	35,5	84,5

Cost + SLA

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,2	22,9	24,6	20,1	13,3	6,7	3,7	1,0	0,3	0,1	0,1	0,0	15,3	94,8	99,8
7,2	8,3	20,0	26,0	20,1	11,7	8,0	4,3	1,1	0,3	0,1	0,1	0,0	15,5	94,1	99,8
10,8	6,0	18,5	23,6	21,0	14,1	8,8	5,7	1,5	0,5	0,2	0,1	0,0	16,8	92,0	99,7
14,4	5,6	16,5	20,9	19,4	15,1	10,5	8,4	2,4	0,7	0,3	0,2	0,1	18,4	88,0	99,5
18,0	5,1	14,5	18,6	17,8	14,5	12,5	11,4	3,5	1,2	0,5	0,2	0,1	20	83,0	99,1
21,6	4,0	11,5	15,7	16,1	15,2	13,9	14,4	5,8	1,9	1,0	0,3	0,3	22,7	76,4	98,5
25,2	3,0	9,2	12,5	14,4	14,5	14,2	17,9	8,4	3,2	1,4	0,6	0,6	25,6	67,8	97,3
28,8	2,4	7,2	10,3	12,9	13,8	14,5	20,2	9,9	4,5	2,2	1,1	1,1	28,3	61,1	95,7
32,4	1,3	4,4	7,1	10,1	11,7	13,8	24,1	12,4	7,0	3,5	2,1	2,3	33,3	48,4	91,9
36,0	0,8	2,9	5,0	7,0	10,0	12,3	23,3	15,5	9,1	5,4	3,0	5,6	39,2	38,0	85,9

Waiting + Cost

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,3	23,0	24,8	20,0	13,3	5,9	3,7	1,1	0,3	0,1	0,3	0,1	15,4	94,3	99,4
7,2	8,3	20,6	26,9	20,1	11,9	6,0	3,6	1,5	0,6	0,3	0,2	0,1	15,5	93,8	99,5
10,8	6,4	18,7	25,4	21,2	13,7	6,4	4,7	1,8	0,7	0,4	0,3	0,2	16,8	91,8	99,0
14,4	5,9	17,4	22,9	20,7	14,0	7,7	6,2	2,8	1,2	0,5	0,3	0,4	18,2	88,6	98,8
18,0	5,5	15,1	20,4	19,6	14,3	9,6	8,9	3,6	1,6	0,6	0,4	0,4	19,7	84,5	98,6
21,6	4,1	13,0	18,0	17,9	15,4	10,1	11,0	5,1	2,3	1,5	0,7	0,9	22,4	78,5	96,9
25,2	3,7	11,5	15,3	17,2	14,2	11,2	12,5	6,4	3,4	2,0	1,1	1,4	24,7	73,1	95,4
28,8	3,0	9,3	14,6	15,3	14,4	11,5	14,4	7,5	4,2	2,2	1,5	2,1	26,9	68,1	94,2
32,4	2,1	7,1	11,2	13,5	13,8	12,0	16,7	9,4	5,7	3,2	2,0	3,3	30,8	59,7	91,5
36,0	1,5	5,6	9,6	11,4	12,0	11,2	16,9	11,2	7,1	4,6	3,1	5,7	35,5	51,3	86,5

FIGURE 7. On line instance objectives and their end of day performance

value of 60 minutes. Experiments however revealed that the 60 minutes threshold performance results are better when we consider a threshold of 30 minutes for the on line instances. (A likely explanation is that a threshold of 60 minutes in the on line optimization might cause any unexpected delay in service times or arrival of new events to entail a service time of more than 60 minutes.) Selecting the threshold performance as the only objective for the on line instance results in poor performance, since it doesn't discriminate between assignments within the threshold bound. Hence we have added two different choices of objectives, yielding two different bicriteria functions for the on line optimization problems.

A first bicriteria function is obtained by optimizing a weighted average of the waiting time and the waiting time in excess of the threshold value. Letting the latter dominate by weighing it with a factor of 7.5, we obtained the second group of results. Figure 8 displays clearly that most categories with short waiting times have less customers, while there is an increased number of customer waiting longer than the threshold value.

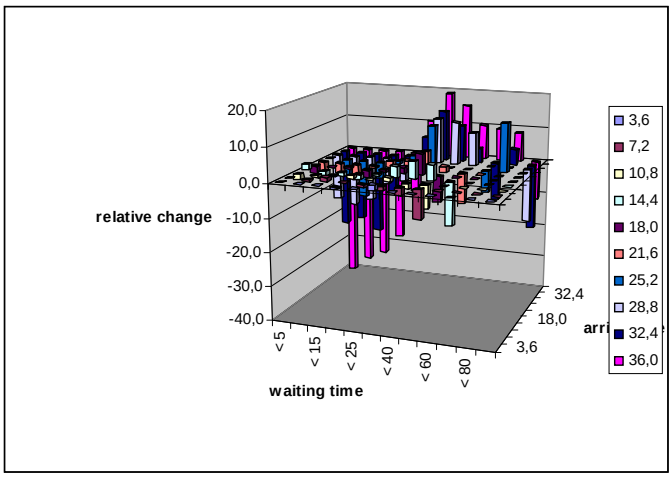

FIGURE 8. Waiting Time + SLA versus Waiting Time on line objectives

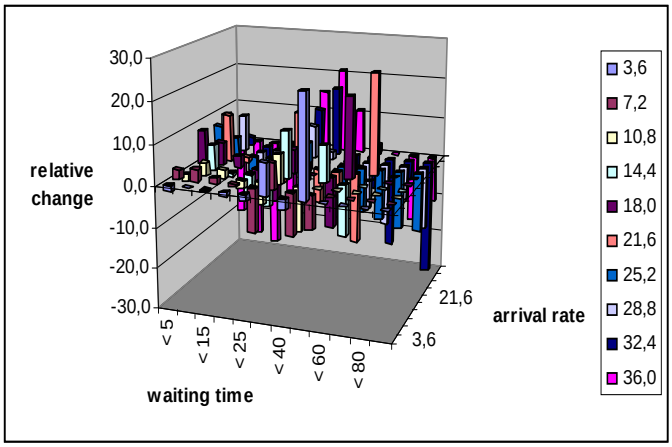

FIGURE 9. Travel Cost + SLA versus Waiting Time

A second bicriteria objective is where we minimize a weighted combination of the travel time and the waiting in excess of the threshold value. Using the same weight of 7.5 we obtain the result of the third group in Figure 7, and the relative performance is displayed in Figure 9. We observe that compared to the base scenario the explicit consideration of waiting time leads to a higher proportion of request being served quickly, and a most reduction in the portion of customers with waiting times in excess of 60 minutes. The number of requests waiting between 30 and 60 minutes has grown and therefore the average waiting time hasn't improved significantly.

As the bicriteria objective of travel time and threshold waiting outperforms the bicriteria objective of waiting time and threshold waiting, one

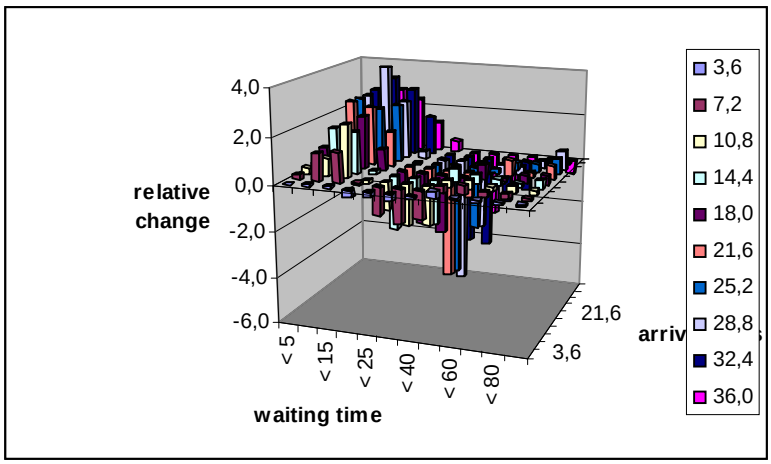

FIGURE 10. Waiting Time + Travel Cost versus Waiting Time

might be interested in optimizing the on line instances using travel cost only. Using solely an objective of travel cost for the on line instances however produces solutions in which some service men serve many customers while others are idle. These solutions result in very poor waiting time performance. Nevertheless, considering a bicriteria objective of the sum of the waiting time and the travel costs produce the fourth group of results in Figure 7, and relative changes as depicted in Figure 10. The results in Figure 10 display consistently more requests in short waiting categories and less requests in longer waiting categories. This is confirmed by the consistent reduction in average waiting time (around 5%) in Figure 7, and the higher number of requests within the threshold values of 30 and 60 minutes. The improvements are especially considerable for the heavy traffic instances which we consider to be more relevant. Minimizing the sum of the waiting time and the travel cost results in 51.3% of requests being served within the threshold of 30 minutes, whereas considering only the waiting time yields a waiting time of 30 minutes or less for only 42.3% of customers. In all other categories, the percentages of customers is reduced and by consequence the end of day average waiting time has dropped from 38.8 to 35.5 minutes.

Completely omitting the penalty for waiting in excess of a threshold entails the risk of excessive long waiting for badly positioned customers (see [Huigenbosch et al. 2008] for an example), which might be considered undesirable as well (as it leads to dissatisfaction and/or churn).For all practical purposes, it is therefore worth considering to let human planners take care of badly positioned customers who might otherwise be left unattended, or to incur indeed a heavy penalty for waiting in excess of a high threshold value (for the instance under consideration, it might be set at 90 minutes).

The results in this section clearly indicate that one must be careful in selecting an objective function to solve the on line instances. It is not trivial to find a multicriteria objective function for the on line instances, that is for the dispatching decisions, which supports the business performance indicators management chooses for end of day performance. Carefully setting end of day performance measures which fit the competitive priorities is therefor a first step in steering the dispatching. A necessary second step is to understand the relationship between these end of day performance objectives and the (multicriteria) objectives available for solving the on line instances. Interestingly, our results indicate that steering operations with SLA threshold based objectives delivers worse end of day threshold performance than steering operations on a combination of classical objectives such as average waiting time and travel time.

7. Diversion

Diversion refers to the undoing of the assignment of a service man to a request after the service man has started traveling towards the request. It is considered by various authors in the context of real time vehicle routing (see e.g. [Ichoua et al. 2000] and the references therein.) For a service operation, even more radical planning changes might be considered. Consider for instance the case where an ambulance is loading a patient from a hospital to transport it to an elderly home for further recovery, when an emergency case arrives. Then preempting the current service operation and shifting to the newly arrived request is certainly an improvement. In real life however, many service companies don 't encounter such urgent requests, and don't consider diversion, i.e. the assignment of a service man after the traveling towards the requests has started. In the experiments in previous sections we have not allowed diversion, but have allowed changing the assignments of planned request, that is the next request of a service man as long as he hasn't started traveling. This is not only advantageous since it allows to deal with newly arriving customer requests, it also allows to more flexibly reassign requests when repair times are shorter or longer than expected.

In practice, diversion is not popular among service men to whom it may give the impression that the planning process functions poorly. The nearer the destination, the more serious the concerns regarding diversion. Hence we have conducted experiments in which diversion is allowed, but only for repair men who are not close yet to their next request

Since diversion particularly allows to divert service men to newly arriving requests, it is reasonable to expect that it increases the relative number of fast responses, perhaps at the cost of longer waiting times for others. In our experiments we solve diversion scenarios using the setting of the base scenario of the set partitioning approach. We compare the results with the corresponding scenario in which diversion is allowed until a service man has

Set Partitioning base

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,3	22,9	24,7	20,3	13,5	6,2	3,8	0,8	0,2	0,0	0,2	0,0	15,2	94,9	99,7
7,2	8,1	19,4	25,6	20,0	13,1	7,5	4,8	1,2	0,2	0,1	0,0	0,0	15,7	93,7	99,9
10,8	6,1	17,9	23,1	21,4	14,9	8,2	6,4	1,5	0,3	0,2	0,0	0,0	16,9	91,6	99,8
14,4	5,6	15,5	21,1	20,5	16,3	9,4	8,1	2,2	0,8	0,2	0,1	0,0	18,3	88,4	99,5
18,0	4,7	13,7	18,1	18,7	15,7	11,8	11,4	4,0	1,0	0,5	0,1	0,0	20,2	82,7	99,1
21,6	3,9	10,3	15,5	16,4	15,9	12,2	15,7	6,2	2,3	0,8	0,5	0,3	23,1	74,2	98,4
25,2	2,8	8,8	13,0	14,7	14,3	13,5	17,3	9,1	3,6	1,6	0,7	0,7	26	67,1	97,1
28,8	2,2	6,6	10,6	12,8	14,1	13,4	19,8	10,4	5,1	2,6	1,2	1,3	29	59,7	95,0
32,4	1,3	4,3	7,8	10,6	12,1	12,4	20,6	13,4	7,6	4,5	2,2	3,2	34,3	48,5	90,1
36,0	1,0	3,9	6,9	9,1	10,7	10,7	19,3	14,0	8,9	5,4	3,6	6,3	38,8	42,3	84,5

Set Partitioning diversion

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,6	23,8	25,1	19,6	13,1	6,0	3,5	1,1	0,1	0,0	0,2	0,0	14,9	95,2	99,9
7,2	8,5	20,7	27,3	21,0	11,4	6,5	3,4	0,9	0,2	0,1	0,0	0,0	14,9	95,4	99,9
10,8	6,7	19,5	24,9	21,8	14,0	6,0	5,0	1,4	0,4	0,2	0,0	0,0	16,1	92,9	99,7
14,4	6,6	18,1	23,5	20,9	13,7	8,2	6,0	1,8	0,7	0,2	0,2	0,1	17	91,0	99,5
18,0	6,1	17,0	22,0	20,0	13,6	8,8	8,2	2,5	1,0	0,4	0,2	0,1	18,2	87,5	99,2
21,6	5,7	15,0	19,7	18,5	14,9	9,7	9,8	3,7	1,7	0,7	0,3	0,3	20	83,5	98,7
25,2	4,9	14,5	18,0	17,9	14,1	9,9	10,4	5,3	2,5	1,2	0,7	0,5	21,6	79,3	97,5
28,8	4,9	13,0	17,0	15,9	13,6	10,4	12,1	6,3	3,0	1,7	0,9	1,2	23,5	74,8	96,2
32,4	4,0	11,5	14,5	15,0	12,6	10,1	13,7	7,7	4,4	2,7	1,5	2,4	26,8	67,7	93,5
36,0	3,8	10,2	12,8	12,4	11,6	9,7	14,0	9,1	5,6	3,7	2,4	4,6	30,9	60,5	89,2

FIGURE 11. Diversion versus Base

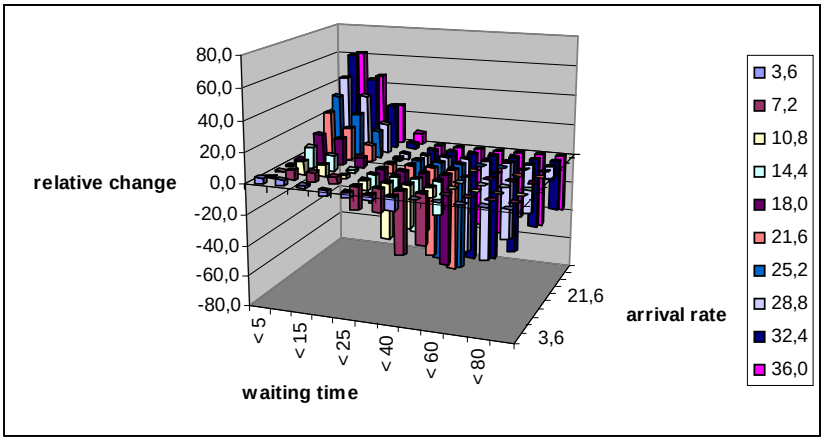

FIGURE 12. Diversion versus Base: Waiting time performance

a remaining travel time of 7.5 minutes. Figure 12, which graphically displays the relative changes reported in Figure 11 clearly displays that even the mild form of diversion introduced in our experiments leads to considerable improvements for the heavy traffic scenarios. There are considerable increases in the number of requests served within 25 minutes, and a significant reductions in the numbers of requests in each of the other categories. The average waiting time for the heavy traffic instances is reduced by more than 20 percent.

Figure 13 presents the differences in total traveled distance and average traveled distance until the next request for various request arrival rates. For the base scenario travel time and distance are equal because of the normalization of travel speed. The results entail a reduction of total traveled

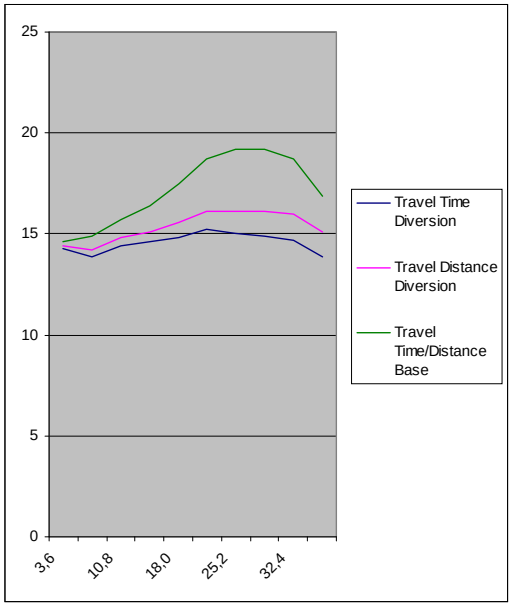

FIGURE 13. Diversion versus Base: travel times and distances

distance by 10.7%, by which the diversion yields a reduction of average travel time to the next request of 17.7%.

For the ASAP, our results clearly indicate that diversion yields an enormous improvement potential over the case where it is disregarded. The improvements obtained are significantly larger than the results reported by Ichoua et al. [2000] on a vehicle routing problem where service duration is zero. The stochasticity of the service duration might therefore be concluded to drive the potential of diversion. Despite the operational disadvantages, we therefore consider it to have huge potential in the quest for improved efficiency and responsiveness for servicing operations where the aim is to service requests as soon as possible.

8. The value of service time information

In the practical application which motivated this work, the service time durations follow an exponential distribution. When basing dispatching decisions on expected waiting times, we therefore use the expected service time in our calculations. In this section we propose two potential improvements. A first scenarios explores the degree of the difficulty caused by the stochasticity of the exponentially distributed service durations. To do so, we consider scenarios where the service time is sampled from the same exponential distribution, but known upon arrival of the request. This might apply to the case where much information on the nature of the requested service is known upon arrival of the request. (An example is the case of servicing

Set Partitioning base

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,3	22,9	24,7	20,3	13,5	6,2	3,8	0,8	0,2	0,0	0,2	0,0	15,2	94,9	99,7
7,2	8,1	19,4	25,6	20,0	13,1	7,5	4,8	1,2	0,2	0,1	0,0	0,0	15,7	93,7	99,9
10,8	6,1	17,9	23,1	21,4	14,9	8,2	6,4	1,5	0,3	0,2	0,0	0,0	16,9	91,6	99,8
14,4	5,6	15,5	21,1	20,5	16,3	9,4	8,1	2,2	0,8	0,2	0,1	0,0	18,3	88,4	99,5
18,0	4,7	13,7	18,1	18,7	15,7	11,8	11,4	4,0	1,0	0,5	0,1	0,0	20,2	82,7	99,1
21,6	3,9	10,3	15,5	16,4	15,9	12,2	15,7	6,2	2,3	0,8	0,5	0,3	23,1	74,2	98,4
25,2	2,8	8,8	13,0	14,7	14,3	13,5	17,3	9,1	3,6	1,6	0,7	0,7	26	67,1	97,1
28,8	2,2	6,6	10,6	12,8	14,1	13,4	19,8	10,4	5,1	2,6	1,2	1,3	29	59,7	95,0
32,4	1,3	4,3	7,8	10,6	12,1	12,4	20,6	13,4	7,6	4,5	2,2	3,2	34,3	48,5	90,1
36,0	1,0	3,9	6,9	9,1	10,7	10,7	19,3	14,0	8,9	5,4	3,6	6,3	38,8	42,3	84,5

Set Partitioning median

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,1	22,6	25,2	20,1	13,8	6,0	3,6	1,0	0,3	0,1	0,1	0,1	15,3	94,8	99,7
7,2	8,2	20,0	26,6	20,3	12,4	6,8	3,9	1,3	0,3	0,1	0,1	0,0	15,5	94,3	99,8
10,8	6,2	18,4	24,5	21,9	14,2	7,4	5,1	1,6	0,5	0,1	0,1	0,0	16,5	92,6	99,8
14,4	5,8	16,6	22,0	20,3	15,0	9,2	7,3	2,4	0,9	0,4	0,1	0,1	18,1	88,9	99,5
18,0	5,0	14,7	19,5	19,6	15,2	10,5	10,3	3,0	1,3	0,5	0,2	0,1	19,6	84,5	99,1
21,6	4,2	11,5	16,9	17,2	15,8	11,5	13,4	5,6	2,1	1,1	0,4	0,3	22,3	77,1	98,2
25,2	3,3	10,4	14,3	16,1	14,8	12,3	14,8	7,5	3,5	1,5	0,8	0,6	24,7	71,2	97,0
28,8	2,5	7,4	12,2	14,0	14,5	12,8	17,9	9,1	4,7	2,2	1,2	1,5	27,9	63,4	95,1
32,4	1,4	5,5	9,4	11,9	13,0	12,6	19,2	11,8	6,6	3,8	2,1	3,0	32,4	53,8	91,4
36,0	1,1	4,3	7,8	9,8	11,2	10,8	18,5	13,5	8,3	5,2	3,6	6,0	37,7	45,0	85,3

Set Partitioning service duration known

	< 5	< 10	< 15	< 20	< 25	< 30	< 40	< 50	< 60	< 70	< 80	> 80	average	within 30	within 60
3,6	7,1	23,4	24,7	19,6	13,8	6,4	3,9	0,9	0,1	0,1	0,1	0,0	15,2	95,0	99,9
7,2	8,0	19,6	26,8	21,0	13,1	6,7	4,0	0,7	0,0	0,0	0,0	0,0	15,2	95,2	99,9
10,8	6,2	18,4	24,7	22,3	14,5	7,4	5,2	1,0	0,1	0,0	0,0	0,0	16,1	93,5	99,8
14,4	5,8	16,4	22,4	20,7	15,4	9,4	7,5	1,7	0,5	0,1	0,0	0,0	17,5	90,1	99,8
18,0	5,0	14,4	20,2	18,8	16,1	11,4	10,0	3,0	0,8	0,2	0,1	0,0	19,1	85,9	99,7
21,6	4,0	11,6	16,6	18,1	16,2	12,5	13,9	5,0	1,4	0,4	0,2	0,0	21,5	79,0	99,3
25,2	3,2	10,0	14,4	16,3	15,4	12,8	16,3	7,3	2,7	0,9	0,4	0,2	23,9	72,1	98,4
28,8	2,4	8,1	12,0	14,8	15,1	12,6	18,4	9,5	4,0	1,7	0,7	0,7	26,7	65,0	96,9
32,4	1,4	5,3	9,1	11,6	13,1	12,6	21,3	12,6	6,5	3,3	1,3	1,7	31,5	53,1	93,5
36,0	1,1	4,3	7,8	9,8	11,3	11,8	20,0	13,7	8,6	4,9	2,6	4,1	36	46,1	88,4

FIGURE 14. Service Time Information variants

machines which are able to communicate there condition using internet or mobile communication. An alternative is to interview the customer.) In any case the experiment provides insight in the potential improvement that can be obtained by acquiring service time information. Another improvement which we considered is not to consider the expected service duration in the calculations, but the median duration. Typically there is a group of candidate service men for a request, and therefore it is more likely to be serviced by a service man who completes his current service earlier than expected, than by a service man who finishes later than expected. In view of the probability density of the exponential distribution (which has a long tail) we therefore report on experiments where the expected waiting times are not based on expected service durations but on median service durations.

The results are presented in Figure 14. We see that even for the case where the service times are a priori exactly known, the improvements are relatively modest.

Figure 15 graphically compares the base case with the case of perfect service time duration. We observe that there are consistently more requests serviced quickly when the durations are known, and consistently fewer requests having waiting times of more than 25 minutes. Hence the end of day average and threshold performances have improved. The percentage of request served within 30 minutes for the heavy traffic instance is 45.1% and

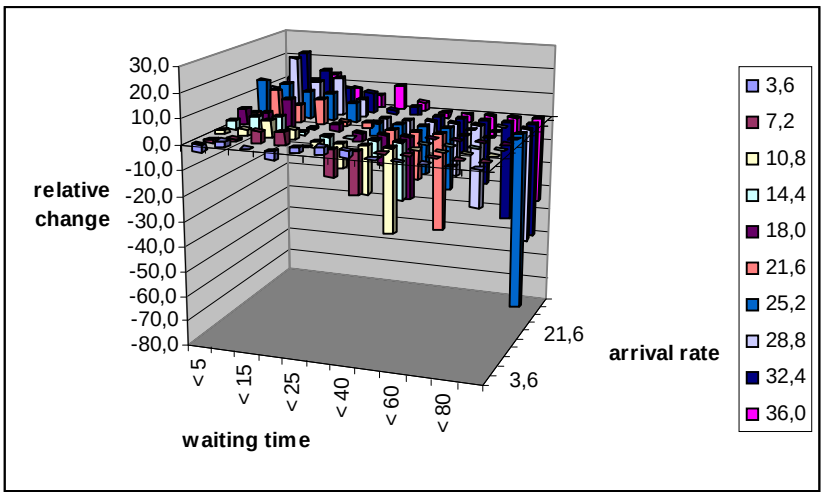

FIGURE 15. A priori known durations versus base scenario

the improvement of the average waiting time is 7.2%. The percentage of request waiting longer than 60 minutes drops to 11.4%.

Interestingly, working with the median instead of the expected value, already realizes a non negligible part of the thus available improvement, as is displayed in Figure 16. We observe again a consistent increase of the numbers of request in categories receiving service within 25 minutes. The service provided to customers waiting longer is not decreased as consistently, but the overall end of day performance appears to have improved. More precisely, taking the median instead of the average results in on time percentage of 44.0%, as opposed to 42.3% and a reduction of the percentage of requests waiting more than 60 minutes of 0.9%. The average waiting time is consistently lower for heavier traffic instances as well.

Knowing service duration exactly a priori is unrealistic for all practical purposes, but by interviewing or technology it is possible to obtain specific information regarding the service requested and therefore of the expected duration..Under the settings of our simulation scenarios, however, knowing the service times a priori only allows a relatively modest improvement of the end of day performance when compared to changing objectives for the on line instances, or allowing diversion. This should lead to caution regarding investments in technology and processes to improve the quality of a priori information of service request. (Of course real life applications exists for which this information is vital, our conclusions regard the settings of this research.) A minor improvement which is for free is however, appears to be attainable by working with the median service duration rather than the expected service duration in the planning.

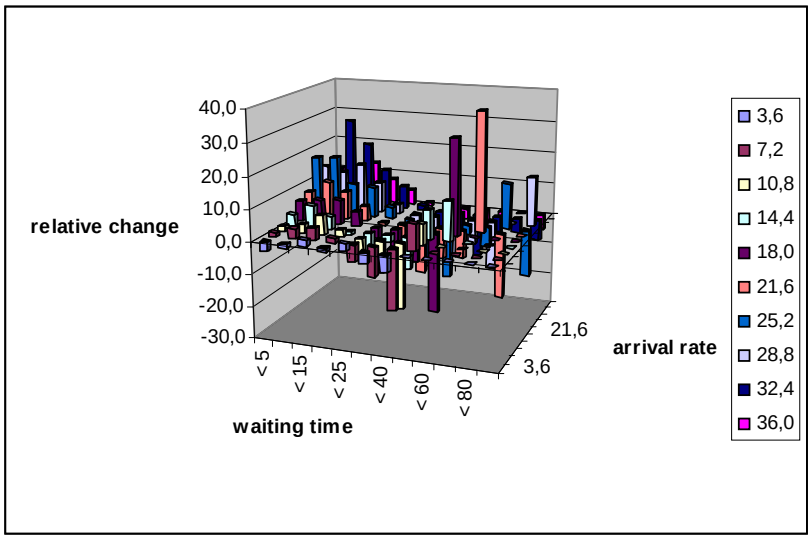

FIGURE 16. Median service times versus base scenario

9. Conclusions and Further Research

In this paper we have analyzed the real time problem of assigning service men to request, which we named ASAP, the after salesman problem. After sales, long term, customer relations continue to gain importance, as long term service based customer relationships are seen as a much more sustainable and value creating business model than manufacturing and marketing of goods. Hence solving the ASAP is relevant for all service organizations, whether from pure service industry, or affiliated with an OEM. In ASAP, customer satisfaction and hence responsiveness to requests is more appropriate as an objective than classical alternatives such as the total travel distance, the sum of the completion times, or the average waiting time. We proposed several objective functions which capture responsiveness, noting that many of the work reported in the literature focusses on efficiency rather than on drivers of customer satisfaction. We advocate models in which service level agreements which specify threshold values on request response times are used. Performing extensive simulation studies, we analyzed several models and solution methods as encountered in practice as well as in scientific literature. The set partitioning model which maintains a planning in which all known requests are planned is shown to perform best especially under the practically relevant heavy traffic scenarios. Of course the simulation scenarios we have analyzed make certain assumptions regarding arrival times and rates, service duration, travel speed, size of the service area, et cetera, and therefore the results are not generally applicable. The scenarios are however

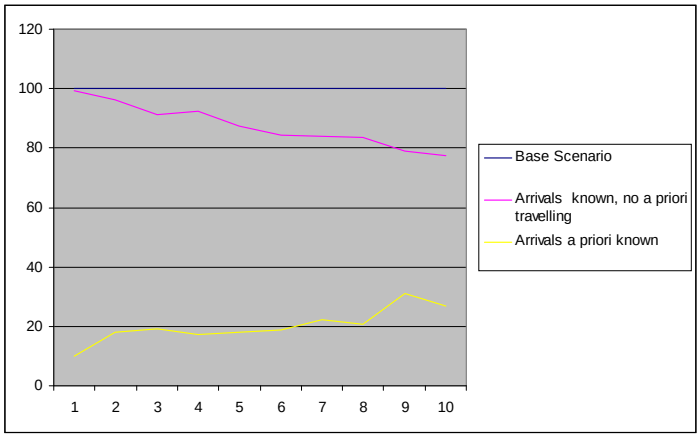

FIGURE 17. Knowing Arrivals A Priori

derived from real life data of the largest service organization in the Netherlands for which the set partitioning approach is implemented and running satisfactorily 24/7, 365 days per year[Huigenbosch et al. 2008].

Although we have argued that objectives such as total travel time or average waiting time are conflicting with customer satisfaction rather than supporting it, our analysis shows that in order to obtain maximum end of day performance with respect to on time arrival, steering the on line assignment on a combination of travel distance and waiting time provides significantly better results than alternatives which do take the thresholds explicitly into account. Although we have an intuitive understanding of these results, which don't only apply to the set partitioning model, but also to the matching model, we feel that a better theoretical understanding of these phenomena is required. The same holds with respect to the length of the reoptimization period. Somewhat counterintuitively, increasing the length is more beneficial as the arrival rate of requests increases. A more fundamental understanding of this relationship is welcomed.

Diversion refers to the possibility to change the destination of a service man who is already on its way to a request. Since diversion is preferably avoided in real life settings we know of, we have disregarded it in our basic experiments. However when allowing diversion as long as a service man is not within 7.5 minutes of the destination (out of an average travel time of around 15 minutes), yields performance results improvements which are bigger than any of the other modifications considered in our computational experiments. Hence we strongly recommend considering practical implementation as well as a further theoretical exploration of its benefits.

Another area of potential large improvements is to reduce the uncertainty in the service durations. We have modelled service durations after real life data to be exponentially distributed, allowing for large deviations from the

expected value. In the experiments, the average service time is 15 minutes, which is close to the realized average travel times. Nevertheless, our experiments show that even knowing the service durations completely as soon as requests arrive results in a end of day performance improvements yields relatively modest improvements. Combined with the results on diversion this leads to the informal conclusion that the real time complexity of the problem is in the arrivals of the requests rather than in their servicing. This is further confirmed by Figure 17 which displays the improvements possible when knowing all requests in advance. It provides two varieties. One in which service men are allowed to travel to requests before they have even arrived, and one on which they can only start travelling after the requests have arrived. (These variations are known under the names of abusive and non abusive adversary scenario's, see e.g. [Grotschel et al. 2001][Krumke et al. 2002b] for related work.). The case where traveling is only allowed after an event has arrived has no clearly defined practical counterpart, since despite the fact that the service man can't travel to it's destination before the request has arrived, it is assumed to be known on beforehand (so why than not travel towards it.) Nevertheless the 25 % improvement in average waiting time is an improvement which is very large. The scenario where traveling in anticipation of requests which haven't arrived yet is allowed, results in a reduction of average waiting time of as much as 80%, confirming indeed that the stochastic arrival process, and not the service durations forms the core of the problem. Hence future work regarding expectations of arrival is called for. Interesting work in this direction has been made for instance in the context of the dynamic traveling salesman problem in transportation oriented applications. Larsen et al. [2004] test various strategies for relocating idle service men. Hvattum et al. [2006,2007] consider strategies where the next requests are selected using scenario based analyses and appear to close a considerable part of the gap between the on line solutions obtained and the off line optimum. Similar ideas are exploited in [Van de Klundert & Otten 2007]. None of these works however apply to the as soon as possible service as requested in the problem studied in this paper, for which such approaches are therefore worthy of attention. Especially the case where requests don't pop up randomly in the plane but only at the locations of customers (as in [Hvattum et al. 2006]) with whom service level agreements are made is in our view a promising and important area.

10. Literature

References

[ANWB 2003] ANWB, 2003, Klanttevredenheidsonderzoek, Kwartaalrapportages 2003, Confidential Publication.

[Huigenbosch et al. 2008] Huigenbosch, P. van, Klundert, J. van de, Wormer, L., 2008, ANWB automates and improves repair men dispatching, Research Memorandum, Faculty of Economics and Business Administration, Maastricht University.

[Bertsimas & van Ryzin 1991] Bertsimas, D. J., Ryzin, G. van, 1991, A stochastic and dynamic vehicle routing problem in the euclidean plane, Operations Research 39(4), 601-615.

[Bertsimas & Van Ryzin 1991b] Bertsimas, D. J., Ryzin, G. van, 1991, Stochastic and dynamic vehicle routing with general demand and inter-arrival time distributions, 1991, Working Paper, Sloan School of Management, MIT,

[Bertsimas & Van Ryzin 1993] Bertsimas, D. J., Ryzin, G.van, 1993, A stochastic and dynamic vehicle routing problem in the euclidan plane with Multiple Capacitated Vehicles, Operations Research 41(1), 60-76.

[Brady & Cronin 2001] Brady, M.K., Cronin, J.J., Some new thoughts on conceptualizing perceived service quality: a hierarchical approach, Journal of Marketing, Vol. 65, 34-49.

[Collier & Wilson 1997] Collier, D.A., Wilson, D.D., 1997, The role of automation and labor in determining customer satisfaction in a telephone repair service process, Decision Science, Volume 28, No. 3, 689-708,

[Davis & Heineke 1998] Davis, M.M., Heineke, J., 1998, How disconfirmation and actual waiting times impact customer satisfaction, International Journal of Service Industry Management, Vol. 9, No 1., 1998.

[Dantzig et al.1954] Dantzig, G., Fulkerson, R., Johnson, S, 1954, Solution of a large-scale traveling salesman problem, Operations Research 2, 393-410.

[Garfinkel 1985] Garfinkel, R.S., 1985, Motivation and Modeling, in The Traveling Salesman Problem, a guided tour of combinatorial optimization, by Lawler, E.D., Lenstra, J.K., Rinnoy Kan, A.H.G., Shmoys, D.B., John Wiley & Sons, pp 17-36.

[Gendreau et al. 1998] Gendreau, M., Potvin, J-Y, 1998, Dynamic Vehicle Routing and Dispatching, Fleet Management and Logistics, pp 115-226, Kluwer Academic Publishers.

[Gronroos 1994] Gronroos, C., 1994, From Scientific Management to Service Management, International Journal of Service Industry Management, Vol 5., No 1, 5-20.

[Grotschel et al. 2001] Grotschel, M, Krumke, S., Rambau, J.,Winter, T., Zimmermann, U., 2001, Combinatorial on line optimization in real time, Technical Report 01-16, Konrad Zuse Zentrum fur Informationstechnik Berlin,

[Gummesson 1987] Gummesson, E., The New Marketing -Developing Long Term Interactive Relationships, 1987, Long Range Planning, Vol. 20, No 4, 10-20.

[Grotschel et al. 1980] Grotschel, M., Lovasz, L., Schrijver, A., 1980, The Ellipsoid Method and its Consequences in Combinatorial Optimization, Report 80-151-OR, University of Bonn.

[Hvattum et al. 2006] Hvattum, L.,M., Lokketangen, A., Laporte, G., 2006, Solving a dynamic and stochastic vehicle routing problem with a sample scenario hedging heuristic, Transportation Science, Vol. 40, No. 4, pp 421-438,

[Hvattum et al. 2007] Hvattum, L.,M., Lokketangen, A., Laporte, G., 2007, A branch and regret heuristic for stochastic and dynamic vehicle routing problems, Networks, pp 330-340,

[Ichoua et al. 2000] Ichoua, Soumia, Gendreau, Michel, Potvin, Jean-Yves, 2000, Diversion Issues in Real-Time Vehicle Dispatching, Transportation Science, Vol 34., No 4., 426-438

[Ichoua et al. 2006] Ichoua, S., Gendreau, M., Potvin, J-Y, 2006, Exploiting knowledge about future demands for real-time vehicle dispatching, Transportation Science, Vol 40., No 2., 211-225

[Irani et al. 2001] Irani, S., Regan, A. C., Lu, X., 2001, On line Algorithms for the dynamic traveling repair problem, Report Institute of Transportation Studies, University of California, Irvine.

[Johnston 1995] Johnston, R., 1995, The determinants of service quality: satisfiers and dissatisfiers, International Journal of Service Industry Management, Vol 5. No. 5, 53-71,

[Kaplan & Norton 1992] Kaplan, R.S., Norton, D.P., The Balanced Scorecard - Measures that drive performance, 1992, Harvard Business Review, Jan-Feb, 71-79.

[Van de Klundert & Otten 2007] Klundert, J., van de, Otten, B., 2007, Models and Heuristics for dynamic revenue optimization in road transport, Research Memorandum, Faculty of Ecnomics & Business Administration, Maastricht University,

[Krumke et al. 2001] Krumke, S. O., Rambau, J., Torres, L. M. 2002, Real-Time Dispatching of Guided and Unguided Automobile Service Units with Soft Time Windows, Lecture Notes in Computer Science, Vol 2461, 417-424.

[Krumke et al. 2002] Krumke, S. O., Rambau, J., Torres, L. M. 2002, Online Dispatching of Automobile Service Units, Report 02-044, ZIB Berlin.

[Krumke et al. 2002b] Krumke, S., Laura, L. Lipmann, M., Marchetti-Spaccamela, A., Paepe, W.E. de, Poensgen, R., Stougie,L, 2002, Non abusiveness helps: an O(1) competitive algorithm for minimizing the maximum flow time in the online traveling salesman problem, Lecture Notes in Computer Science, Vol 2462, 200-214,

[Larsen et al. 2004] Larsen, A., Madsen, O.B.G., Solomon, 2004, M.M., The a priori dynamic traveling salesman problem with time windows, Transportation Science Vol. 38, No 4., 459-472,

[Lu et al. 2002] Lu, X., Regan, A.,. C., Irani, S., 2002, An asymptotically optimal algorithm for the dynamic traveling repair problem, Report Institute of Transportation Studies, University of California, Irvine.

[Parasuraman et al. 1985] Parasuraman, A., Zeithaml, V.A., Berry, L.L., 1985, A conceptual model of service quality and its implications for future research, Journal of Marketing, 49, 41-50

[Psaraftis 1995] Psaraftis, H.N., 1995, Dynamic Vehicle Routing: Status and Prospects, Annals of Operations Research 61, 143-164.

[Sitters et al. 2003] Sitters, R. A., Stougie, L., Paepe, Willem E. de, 2003, A competitive algorithm for the general 2-server problem, Report, Technische Universiteit Eindhoven.

[Wormer 2005] Wormer, L., 2005, On-line toewijzen van autoreparatiediensten, Master Thesis (In Dutch), Maastricht University.

[Zeithaml et al. 1993] Zeithaml, V.A., Berry, L.L., Parasuraman, A, 1993, The nature and determinants of customer expectations of service, Journal of the academy of Marketing Science, Vol. 21., No 1, 1-12.